LAB ASSIGNMENT - 01

Name - Aniket Kumar

Roll no:- 13

ERP - 1032171203

Batch :- 1

Problem Statement :-

Program to read a paragraph from a text file. Print the paragraph after removing the stop words. Identify part of speech of each words in the paragraph. [Stremming, POS tagging, bag of words and their frequency, finding all meaningful words from given words, listing n-grams] Use NLTK.

Objectives:-

1. To study and explore NLTK for text Processing
2. To learn concepts text processing in NLP.

Theory :-

Explain following concepts.

I. Text processing Concepts.

1] Tokenization :-

In natural language processing, tokenization is the text processing task of breaking up text into smaller components of text (known as tokens)

Stemming:-

Stemming is the text preprocessing normaliyat-ion task concerned with bluntly remaing words affexes (prefixes and suffixes)

lemmatiyation:-

lemmatiyation is the text preprocessing normaliyation task concerned with bringing words down to their root forms.

Ex:- tokenized = {'so", "many", "squids", "are", "jumping"}

Stemmed = {'so', 'many', 'squid', 'be', 'jump']

PoS Tagging:-

In natural language processing, part-of-speech tagging is the process of assigning a part of speech to every word in a string. Using the part of speech can improve the result of lemmatiyation.

Stop word Removal:-

Stop word removal is the process of removing words from a string that dont provide any informat-ion about the tone of a statement.

II Bag of words (BOW)

a) Bag of words is a Natural language processing technique of text modeling.

b) A bag of words is a representation of text that describes the occurrence of words within a document.

c) We just keep track of word counts and disregards the grammatical details and the word order.

d) It is called a "bag" of words because any information about the order or structure of words in the document is discarded

e) The model is only concerned with whether known words occur in the document, not where in the document.

N-grams

An N-gram is an N-token sequence of words: a 2-gram (more commonly called a bigram) is a two-word sequence of words like "really good", "not good", or "your homework", and a 3-gram (more commonly called a trigram) is a three-word sequence of words like "not at all", or "turn off lights".

for example, the bigrams in the first line of text in the previous section: "This is not good at all" are as follows:-

→ "This is"
"is not"
"not good"
"good at"
"at all"

III NLTK Modules for text processing.

a) Corpora :- a package containing modules of example text.

b) tokenize :- functions to separate text strings.

c) Probability :- for modeling frequency distributions and probabilistic System.

d) stem :- package of functions to stem words of text.

e) wordnet :- interface to the WordNet lexical resource

f) chunk :- identify short non-nested phrases in text.

g) etree :- for hierarchical structure over text.

h) tag :- tagging each words with part-of-speech, sense, etc.

i) parse :- building trees over text - recursive descent, shift-reduce, probabilistic, etc.

j) cluster :- clustering algorithms.

k) draw :- visualize NLP structures and processes.

1.) Contrib :- various pieces of software from outside contributors.

Paltform :- 64- bit Open Source Linux, Jupyter Notebook.

Input :- Any text/doc file containing text paragraph in English language.

Output :- Tokens, Text after removing stop words, Tokens with POS tagging, Stem form of text.

Conclusion :- Hence, learned the concepts of text processing in NLP and implemented using NLTK Library.

FAQ's :-

1) Explain the difference between stemming and lemmatization.

Ans. → a) Stemming and lemmatization both generates the root form of the inflected words.

b) Stemming follows an algorithm with steps to perform on the words which makes it faster.

c) whereas, in lemmatization, you used Word Net Corpus and a corpus for stop words as well to produce lemma which makes it slower than stemming.

2) What is Semantic and Syntactic analysis in NLP?

Ans:- Syntactic analysis (Syntax) and Semantic analysis (Semantic) are the two primary techniques that lead to the understanding of natural language.

b) Syntax is the grammatical structure of the text, whereas Semantics is the meaning being conveyed.

c) Syntactic analysis, is the process of analyzing natural language with the rules of a formal grammar

d) Syntactic analysis basically assigns a Semantic structure to text.

e) Semantic analysis is the process of understanding the meaning and interpretation of words, signs and sentence structure

f) Speech recognition, for example, has gotten very good and works almost flawlessly, but we still lack this kind of proficiency in natural language understanding.

Algorithm:-

1. Read a text file in Python using read and open function
2. Tokenize the file into sentences
3. Tokenize each sentence in words and punctuations
4. Remove all the stopwords ('a', 'an', 'the', 'to', & much more)
5. Tag each word to indicate its part of speech.